Submitted to 48th Air Traffic Control Association Annual Conference Proceedings, October 2003

Applying Pairwise Hypothesis Testing In Trajectory Accuracy Analysis

Mike M. Paglione and Lori Charles
Federal Aviation Administration, William J. Hughes Technical Center
Atlantic City Int'l Airport, New Jersey 08405

***Mike M. Paglione** is the Conflict Probe Assessment Team Lead in the Simulation & Analysis Group with the Federal Aviation Administration. **Lori Charles** is a Systems Engineer with Veridian Corporation.*

Abstract
Trajectory accuracy plays an important role in evaluating air traffic management decision support tools. To increase the confidence in trajectory accuracy, various statistical approaches have been applied and not all are sufficient in handling data that has multivariate characteristics. The Federal Aviation Administration's Conflict Probe Assessment Team (CPAT) has developed a practical methodology using a paired-data hypothesis test. The technique properly blocks out nuisance factors and focuses the analysis on the factor under study. This is very useful when air traffic data is very heterogeneous, which is often the case. For example, a given traffic sample will have many flights with various aircraft types, following different routes and altitude profiles, resulting in substantially different accuracy performance. Another practical benefit of the technique is the capability of ranking the individual accuracy performance of a given set of flights against a baseline of performance. As a result, the approach supports regression testing as well as overall system measurement.

Introduction

To achieve the goals of Free Flight, broad categories of advances in ground and airborne automation are required. The Federal Aviation Administration (FAA) has sponsored the development of several ground based air traffic management decision support tools (DSTs) to support the en route and terminal air traffic controllers. A fundamental component of a DST's design is the trajectory modeler, upon which its functionality is based. The trajectory modeler provides a prediction of the aircraft's anticipated flight path, determined from sources such as the flight plan and radar track data received from the National Airspace System (NAS) Host Computer System (HCS). Therefore, the trajectory accuracy, or the deviation between the predicted trajectory and the actual path of the aircraft, has a direct effect on the overall accuracy of these automation tools.

The Conflict Probe Assessment Team (CPAT) at the FAA's William J. Hughes Technical Center developed a generic method of sampling a set of aircraft trajectories for accuracy measurements, called interval-based sampling. It was defined in [1] and applied in [2] on two of the FAA's most advanced trajectory prediction tools, the NASA-developed Center-TRACON Automation System (CTAS) and the MITRE/CAASD-developed User Request Evaluation Tool (URET) prototype DSTs. Both these systems have since been deployed as

production systems into the NAS. However as systems like these are upgraded over time for new aircraft types and/or new functionalities, there is a need for testing whether the upgrades have not inadvertently introduced inaccuracies in the trajectory modeling function. This type of testing is often referred to regression testing in the software community. Using established inferential statistical techniques, this paper applies a practical method that allows the analyst to state with confidence that the trajectory accuracy was not degraded or whether it was. First a common statistical approach will be presented, next its flaws will be discussed, and finally a recommended technique will be described that addresses these flaws.

Common Methodology

The regression test requires a baseline version of the trajectory modeler software to be run with a given traffic sample. This same traffic sample is then run through the upgraded software, which is referred to as the new release version. Both runs are then processed for trajectory accuracy using the interval-based sampling method as described in [1] and [2]. There are several trajectory accuracy metrics that are normally examined using this process, but for simplicity this paper will focus on the horizontal error.

The distance between the sampled aircraft surveillance position and the time coincident trajectory predicted position is defined as the horizontal error. It is unsigned and measured in units of nautical miles. To compare the baseline against the new release run, the difference between the baseline sample mean and the new release sample mean is calculated. Since the sample mean is a statistic and thus a random variable of the true population mean, a statistical hypothesis test is used that considers the variation in both sample means. If the true population means were known, the difference between the baseline run's mean and the new release run's mean could be calculated exactly. If this difference was not equal to zero, it would be concluded that the runs were not equivalent. As described by Devore in [3], the Two-Sample *t* test provides a statistical hypothesis test that provides a criterion to reject the hypothesis that the sample means (a sample statistic of the true population mean) are not equal. From [3], this null hypothesis is expressed in the following Equation 1.

$$H_o : \mu_b - \mu_n = 0 \qquad \textbf{Equation 1}$$

where μ_b is the population mean of the baseline run and μ_n is the population mean of the new release run.

The alternative hypothesis is the difference in populations means are not equal to zero. The following test statistic is presented in [3].

$$\text{Test statistic} : t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_b^2}{m} - \frac{s_n^2}{n}}} \qquad \textbf{Equation 2}$$

where $\bar{x}$ is the sample mean of the baseline run and $\bar{y}$ is the sample mean of the new release run, s_b^2 is the sample variance of the baseline run and m is the sample size of the baseline run, and s_n^2 and n are the same for the new release run, respectively.

The rejection region of the Two-Sample *t* Test is expressed in the following:

$$\text{Reject null hypothesis if } t \geq t_{\alpha/2,\upsilon} \text{ or } t \leq -t_{\alpha/2,\upsilon} \qquad \textbf{Equation 3}$$

where $t_{\alpha/2,\upsilon}$ or $-t_{\alpha/2,\upsilon}$ are parameters taken from the student-t distribution, α is the significance level of the test, and υ is the degrees of freedom for this test[1]. The test assumes the trajectory measurements from each run are normally distributed random variables, and the runs are independent from one another. These assumptions can be tested, but only the later will be discussed in this paper.

Example Application of Two-Sample *t* Test

To test the hypothesis defined in Equation 1 for the measurements of trajectory horizontal error, two runs were performed on a NAS trajectory modeler and the horizontal error was measured at a look-ahead time of five minutes. The sample scenario was based on two-hours of recorded traffic data from Indianapolis en route center in May 1999. The trajectory modeler produced over 5000 trajectories for each of the runs. The baseline run produced a sample mean of 3.45 nautical miles of horizontal error and a sample standard deviation of 9.62 nautical miles (square root of the sample variance). The new release run produced a sample mean of 3.92 nautical miles and sample standard deviation of 9.14 nautical miles. Since the same traffic sample was run through the trajectory modeler, both runs are balanced with the same number of sample horizontal error measurements of 2347.

By applying Equation 2 on the above values, the test statistic *t* equals –1.72. The rejection region from Equation 3 equals ± 1.96, using a significance of 0.05 and 4680 degrees of freedom. This value is found in Table A.5 from [3] as the critical value taken from a student t distribution. Therefore, the hypothesis that the mean horizontal error of the two runs is equivalent cannot be rejected (i.e. t is not $\geq t_{0.025,4680}$ or $\leq -t_{0.025,4680}$). Therefore, the upgrade or new release trajectory model is considered equivalent to the previous baseline version. Even though the difference in sample means was -0.47 nautical miles, the difference was not great enough to compensate for the variability in taking sample measurements from each population of trajectory error measurements.

[1] This degrees of freedom parameter is a function of the number of samples taken for the test and approximately equal to *m+n-2*. The actual formula is defined in Section 9.2 of [3].

Recommended Methodology

The results in the methodology and example presented above contain significant flaws. The problem lies in the assumption that the two samples are independent. Since the same air traffic sample is input into both runs of the trajectory model, the other variables that influence trajectory accuracy are expressed in the variability of flights in the two runs. These flights are the same for each run, so their influence has a proportional effect on both runs. In other words, if a specific flight exhibits higher than normal error in the baseline run, it would be expected that the same flight would have similar high error in the new release run. Of course some flights may exhibit better performance in the new release, if indeed the upgrade was to reduce these errors, but on average if the flights perform in this manner, it can be said that the runs are not independent. A statistic that measures degree of linear dependence between samples is the correlation coefficient[2]. For the above example, the correlation coefficient was 0.98. This indicates that there is a strong linear dependence between the two runs. An alternative technique is then required.

Fortunately, there is a simple solution known as the Paired *t* Test. Instead of taking the difference between the sample means, the sample measurements are paired for the same flight and position. The large variability between flights and linear dependence between runs is effectively blocked out of the experiment. Taking the difference between paired trajectory measurements of same flight and position from the two runs produces a new statistic, the sample differences. This is expressed in the following equation:

$$D_i = x_i - y_i \qquad \textbf{Equation 4}$$

where *i* is the particular measurement from the two runs, x_i is the trajectory measurement for the baseline run and y_i is the same for the new release run.

Therefore, the hypothesis now is that the sample mean of D_i's is equal to zero. The mean of the difference between two numbers is equal to the difference between the means of the same set of numbers. Referring to the difference between sample means in Equation 2, the $\bar{x} - \bar{y}$ is equal to the sample mean of D_i's or $\bar{d}$. Therefore, while the hypothesis in Equation 1 is the same, the denominator in the test statistic is not. The following equation expresses the Paired *t* Test's test statistic:

$$t = \frac{\bar{d}}{s_D / \sqrt{n}} \qquad \textbf{Equation 5}$$

where the s_D is the sample standard deviation of the differences (i.e. the D_i's) and the n is the sample size of these differences.

[2] The correlation coefficient is explained in detail Section 12.5 in [3]. It is a value between negative one and one. A value close to positive or negative one indicates a strong linear dependence. A value close to zero indicates independence.

The rejection region of the Paired *t* Test is expressed in the following:

$$\text{Reject null hypothesis if } t \geq t_{\alpha/2,n-1} \text{ or } t \leq -t_{\alpha/2,n-1} \qquad \textbf{Equation 6}$$

Example Application of Paired *t* Test

Now repeating the same example as before but applying the Paired *t* Test, the sample mean of the differences is -0.47 nautical miles. The sample standard deviation of the differences is 1.89 nautical miles and the number of differences is 2347. The test statistic is –12 and the rejection region is once again ± 1.96. Therefore, the hypothesis that the mean of the paired differences of horizontal errors between runs is not equivalent and can be rejected (i.e. t is $\leq -t_{0.025,2346}$). This is obviously a very different conclusion than in the previous example. Notice the standard deviation of differences, 1.89, is four to five times smaller than the standard deviations of the trajectory errors for each run. The loss in degrees of freedom (about half) is more than compensated for by the reduction in variance of the samples. As discussed in [3], [4], and [5], the Paired *t* Test has a property of improving the precision of the test statistic when there is a correlation between runs and significant heterogeneity between samples (in this example the difference between flights). Furthermore, the D_i's are easily sorted and flights with the largest differences can be examined in detail providing a very practical method to investigating the errors.

Conclusion

In conclusion, the development and later maintenance of the trajectory modeling function of FAA decision support tools requires frequent regression testing between baseline and new releases of the software. To perform this testing effectively, it is recommended that the Paired *t* Test be used, which has the property of improved precision by reducing the variance in the samples. This allows the runs to be correlated, but still requires the samples to be normally distributed. A future publication will present non-parametric techniques that do not require this assumption to be met.

References

[1] Cale, M, Liu, S, Oaks, R, Paglione, M, Ryan, H, Summerill, S. (December 2001), "A Generic Sampling Technique for Measuring Aircraft Trajectory Prediction Accuracy," Presented at the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM.

[2] Paglione et al. (May 1999), "Trajectory Prediction Accuracy Report: URET/CTAS," (DOT/FAA/CT-TN99/10), WJHTC/ACT-250.

[3] Devore, Jay L. (2000), *Probability and Statistics for Engineering and the Sciences, 5th Edition*.

[4] Montgomery, Douglas C. (1997), *Design and Analysis of Experiments, 4th Edition*.

[5] Paglione, Mike M., (July 24, 2000) "Quick Example of Benefits of Pairing Trajectory Accuracy Data," Presented at Lockheed Martin's URET System Engineering Meeting.